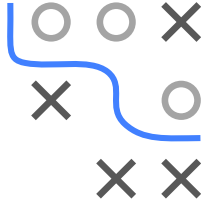
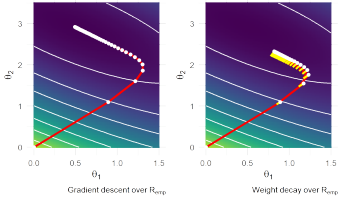
Introduction to Machine Learning



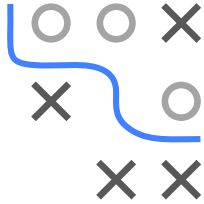
Geometric Analysis of L2 Regularization and Weight Decay



Learning goals

- Have a geometric understanding of *L2* regularization
- Understand why *L2* regularization in combination with gradient descent is called weight decay

WEIGHT DECAY VS. L2 REGULARIZATION



Let us optimize the *L2*-regularized risk of a model $f(\mathbf{x} \mid \boldsymbol{\theta})$

$$\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

by gradient descent. The gradient is

$$\nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}.$$

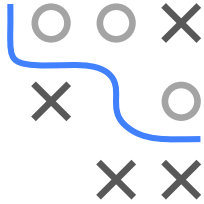
We iteratively update $\boldsymbol{\theta}$ by step size α times the negative gradient

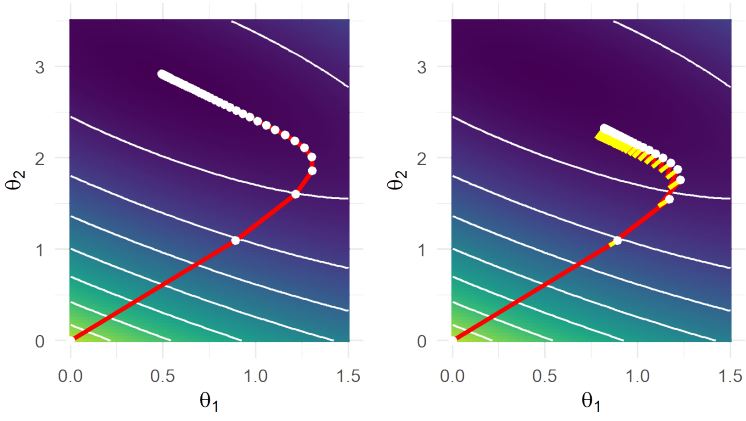
$$\begin{aligned} \boldsymbol{\theta}^{[\text{new}]} &= \boldsymbol{\theta}^{[\text{old}]} - \alpha \left(\nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[\text{old}]}) + \lambda \boldsymbol{\theta}^{[\text{old}]} \right) \\ &= \boldsymbol{\theta}^{[\text{old}]}(1 - \alpha\lambda) - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[\text{old}]}). \end{aligned}$$

The term $\lambda \boldsymbol{\theta}^{[old]}$ causes the parameter (**weight**) to **decay** in proportion to its size. This is a very well-known technique in deep learning - and simply *L2* regularization in disguise.

WEIGHT DECAY VS. L2 REGULARIZATION

When we use weight decay, we follow the steepest slope of $\mathcal{R}_{\text{emp}}$ as for gradient descent, but in every step, we are pulled back to the origin.

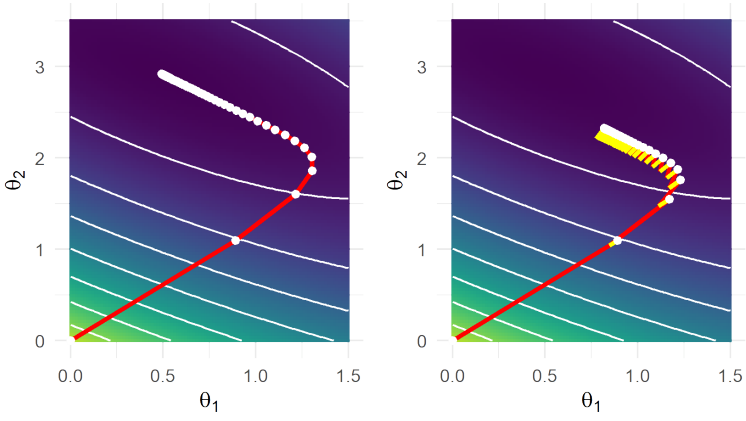




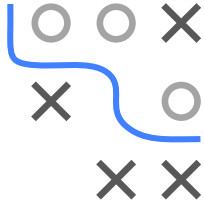
Gradient descent over R_{emp} Weight decay over R_{emp}

WEIGHT DECAY VS. L2 REGULARIZATION

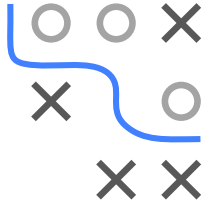
How strongly we are pulled back to the origin for a fixed stepsize α depends only on λ (as long as the procedure converges):



Gradient descent over R_{emp} Weight decay over R_{emp}



WEIGHT DECAY VS. L2 REGULARIZATION



Weight decay can be interpreted **geometrically**.

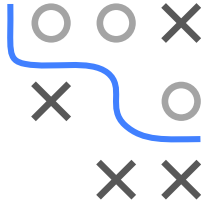
Let's use a quadratic Taylor approximation of the unregularized objective $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ in the neighborhood of its minimizer $\hat{\theta}$,

$$\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\hat{\theta}) + \nabla_{\boldsymbol{\theta}}\mathcal{R}_{\text{emp}}(\hat{\theta}) \cdot (\boldsymbol{\theta} - \hat{\theta}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\theta})^T \boldsymbol{H}(\boldsymbol{\theta} - \hat{\theta}),$$

where $\boldsymbol{H}$ is the Hessian matrix of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ evaluated at $\hat{\theta}$.

- The first-order term is 0 in the expression above because the gradient is 0 at the minimizer.
- $\boldsymbol{H}$ is positive semidefinite.

WEIGHT DECAY VS. L2 REGULARIZATION



The minimum of $\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta})$ occurs where $\nabla_{\boldsymbol{\theta}} \tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta} - \hat{\theta})$ is 0. Now we *L2*-regularize $\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta})$, such that

$$\tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta}) = \tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2$$

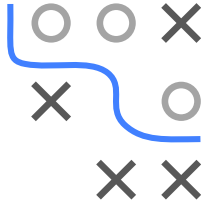
and solve this approximation of $\mathcal{R}_{\text{reg}}$ for the minimizer $\hat{\boldsymbol{\theta}}_{\text{ridge}}$:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta}) &= 0, \\ \lambda \boldsymbol{\theta} + \boldsymbol{H}(\boldsymbol{\theta} - \hat{\theta}) &= 0, \\ (\boldsymbol{H} + \lambda \boldsymbol{I})\boldsymbol{\theta} &= \boldsymbol{H}\hat{\theta}, \\ \hat{\boldsymbol{\theta}}_{\text{ridge}} &= (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \boldsymbol{H}\hat{\theta}, \end{aligned}$$

This gives us a formula to see how the minimizer of the *L2*-regularized version is a transformation of the minimizer of the unpenalized version.

Caveat: Equivalence of weight decay and *L2* regularization only holds for vanilla SGD (not e.g. Adam)

WEIGHT DECAY VS. L2 REGULARIZATION



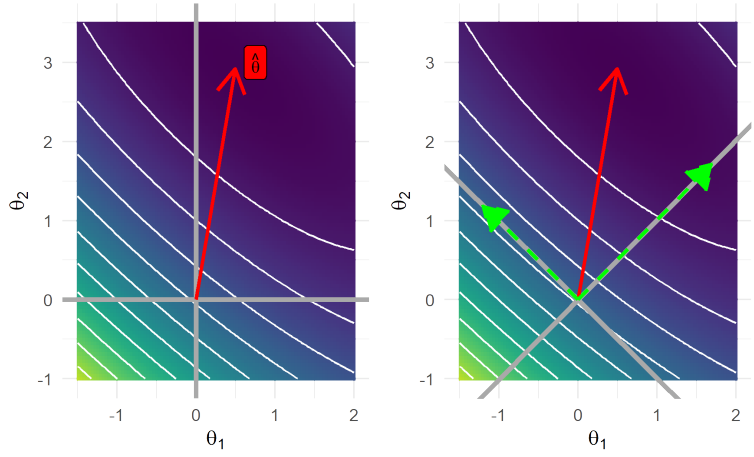
- As λ approaches 0, the regularized solution $\hat{\boldsymbol{\theta}}_{\text{ridge}}$ approaches $\hat{\theta}$. What happens as λ grows?
- Because $\boldsymbol{H}$ is a real symmetric matrix, it can be decomposed as $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top$, where $\boldsymbol{\Sigma}$ is a diagonal matrix of eigenvalues and $\boldsymbol{Q}$ is an orthonormal basis of eigenvectors.
- Rewriting the transformation formula with this:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{ridge}} &= \left(\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\hat{\theta} \\ &= \left[\boldsymbol{Q}(\boldsymbol{\Sigma} + \lambda\boldsymbol{I})\boldsymbol{Q}^\top\right]^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\hat{\theta} \\ &= \boldsymbol{Q}(\boldsymbol{\Sigma} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\hat{\theta}\end{aligned}$$

- Therefore, weight decay rescales $\hat{\theta}$ along the axes defined by the eigenvectors of $\boldsymbol{H}$. The component of $\hat{\theta}$ that is aligned with the j-th eigenvector of $\boldsymbol{H}$ is rescaled by a factor of $\frac{\sigma_j}{\sigma_j+\lambda}$, where σ_j is the corresponding eigenvalue.

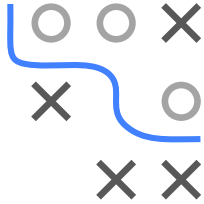
WEIGHT DECAY VS. L2 REGULARIZATION

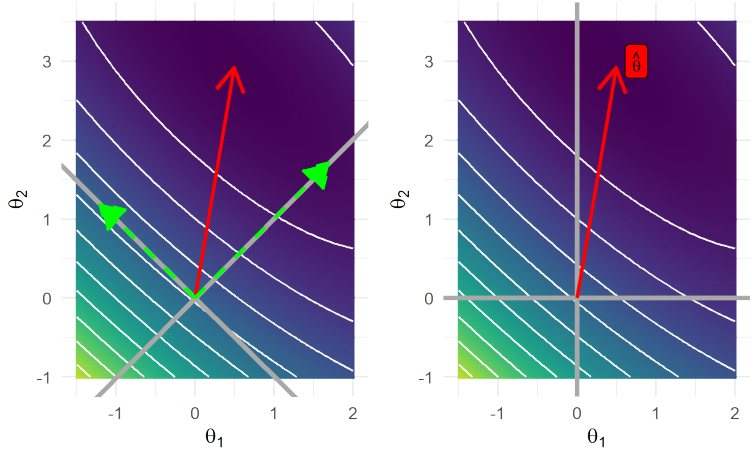
Firstly, $\hat{\theta}$ is rotated by $\boldsymbol{Q}^\top$, which we can interpret as a projection of $\hat{\theta}$ on the rotated coordinate system defined by the principal directions of $\boldsymbol{H}$:



WEIGHT DECAY VS. L2 REGULARIZATION

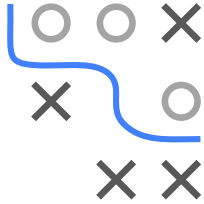
Since, for $\lambda = 0$, the transformation matrix $(\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} = \boldsymbol{I}$, we simply arrive at $\hat{\theta}$ again after projecting back.

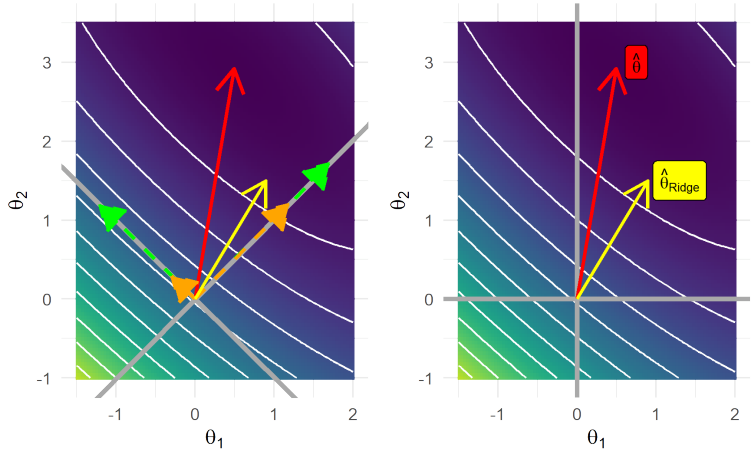




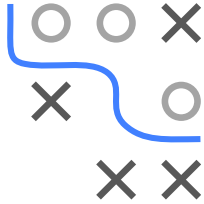
WEIGHT DECAY VS. L2 REGULARIZATION

If $\lambda > 0$, the component projected on the j-th axis gets rescaled by $\frac{\sigma_j}{\sigma_j + \lambda}$ before $\hat{\theta}_{\text{ridge}}$ is rotated back.



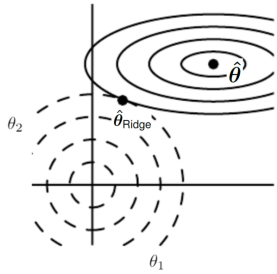


WEIGHT DECAY VS. L2 REGULARIZATION



- Along directions where the eigenvalues of $\boldsymbol{H}$ are relatively large, for example, where $\sigma_j >> \lambda$, the effect of regularization is quite small.
- On the other hand, components with $\sigma_j << \lambda$ will be shrunk to have nearly zero magnitude.
- In other words, only directions along which the parameters contribute significantly to reducing the objective function are preserved relatively intact.
- In the other directions, a small eigenvalue of the Hessian means that moving in this direction will not significantly increase the gradient. For such unimportant directions, the corresponding components of $\boldsymbol{\theta}$ are decayed away.

WEIGHT DECAY VS. L2 REGULARIZATION



Credit: Goodfellow et al. (2016), ch. 7

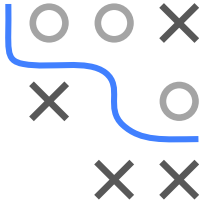
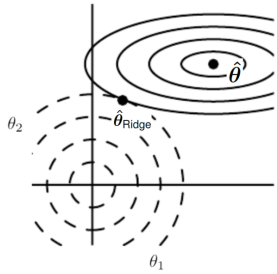


Figure: The solid ellipses represent the contours of the unregularized objective and the dashed circles represent the contours of the *L*2 penalty. At $\hat{\boldsymbol{\theta}}_{\text{ridge}}$, the competing objectives reach an equilibrium.

In the first dimension, the eigenvalue of the Hessian of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ is small. The objective function does not increase much when moving horizontally away from $\hat{\theta}$. Therefore, the regularizer has a strong effect on this axis and θ_1 is pulled close to zero.

WEIGHT DECAY VS. L2 REGULARIZATION



Credit: Goodfellow et al. (2016), ch. 7

Figure: The solid ellipses represent the contours of the unregularized objective and the dashed circles represent the contours of the *L2* penalty. At $\hat{\boldsymbol{\theta}}_{\text{ridge}}$, the competing objectives reach an equilibrium.

In the second dimension, the corresponding eigenvalue is large indicating high curvature. The objective function is very sensitive to movement along this axis and, as a result, the position of θ_2 is less affected by the regularization.